PCT/JP2004/009341

02.7.2004

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年 ７月 １日

出願番号
Application Number: 特願２００３－１８９７１６
[ST. 10/C]: [ＪＰ２００３－１８９７１６]

出願人
Applicant(s): 株式会社山武

２００４年 ８月１２日

特許庁長官
Commissioner,
Japan Patent Office

小川洋

出証番号 出証特２００４－３０７１９７５

【書類名】　特許願
【整理番号】　20030113
【提出日】　平成15年 7月 1日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/20
【発明者】
　【住所又は居所】　東京都渋谷区渋谷２丁目１２番１９号　株式会社　山武内
　【氏名】　村上　英治
【発明者】
　【住所又は居所】　東京都杉並区和泉３丁目１５番１６号
　【氏名】　寺野　隆雄
【特許出願人】
　【識別番号】　000006666
　【氏名又は名称】　株式会社　山武
【代理人】
　【識別番号】　100064621
　【弁理士】
　【氏名又は名称】　山川　政樹
　【電話番号】　03-3580-0961
【手数料の表示】
　【予納台帳番号】　006194
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9722147
【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文章分類装置および方法

【特許請求の範囲】

【請求項１】　文章集合に含まれる複数の文章を、１つ以上の単語からなるタームを複数有するタームリストに基づき分類する文章分類装置において、

前記各文章と前記各タームとの関係を２次元表現したＤＴマトリクスを生成するＤＴマトリクス生成手段と、

グラフ理論で用いられるＤＭ分解法に基づいて、前記ＤＴマトリクス生成手段で得られたＤＴマトリクスを変形することにより、変形ＤＴマトリクスを生成するＤＴマトリクス変形手段と、

このＤＴマトリクス変形手段で得られた変形ＤＴマトリクス上でブロック化されたクラスタごとに、当該クラスタに属する文章を同一分類として出力する文章分類手段とを備えることを特徴とする文章分類装置。

【請求項２】　請求項１において、

任意の前記クラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するラベル生成手段をさらに備えることを特徴とする文章分類装置。

【請求項３】　請求項１において、

前記変形ＤＴマトリクスでの文章の並び順序に応じて、任意の前記クラスタに属する文章またはすべての文章を順に出力する文章編成手段をさらに備えることを特徴とする文章分類装置。

【請求項４】　請求項１において、

任意の前記文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力する要約作成手段をさらに備えることを特徴とする文章分類装置。

【請求項５】　請求項１において、

前記タームリストに対して任意のタームを追加または削除するタームリスト編集手段と、

このタームリスト編集手段による編集前後のタームリストを用いて前記ＤＴマ

トリクス生成手段によりそれぞれＤＴマトリクスを生成し、これらＤＴマトリクスから当該編集の有用性を示す指標を生成して出力する指標生成手段とをさらに備えることを特徴とする文章分類装置。

【請求項６】　文章集合に含まれる複数の文章を、１つ以上の単語からなるタームを複数有するタームリストに基づき分類する文章分類装置で用いられる文章分類方法において、

前記各文章と前記各タームとの関係を２次元表現したＤＴマトリクスを生成するステップと、

グラフ理論で用いられるＤＭ分解法に基づいて前記ＤＴマトリクスを変形することにより、変形ＤＴマトリクスを生成するステップと、

前記変形ＤＴマトリクス上でブロック化されたクラスタごとに、当該クラスタに属する文章を同一分類として出力するステップとを備えることを特徴とする文章分類方法。

【請求項７】　請求項６において、

任意の前記クラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するステップをさらに備えることを特徴とする文章分類方法。

【請求項８】　請求項６において、

前記変形ＤＴマトリクスでの文章の並び順序に応じて、任意の前記クラスタに属する文章またはすべての文章を順に出力するステップをさらに備えることを特徴とする文章分類方法。

【請求項９】　請求項６において、

任意の前記文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力するステップをさらに備えることを特徴とする文章分類方法。

【請求項１０】　請求項６において、

前記タームリストに対して任意のタームを追加または削除するステップと、

編集前後のタームリストを用いてそれぞれＤＴマトリクスを生成し、これらＤＴマトリクスから当該編集の有用性を示す指標を生成して出力するステップとを

さらに備えることを特徴とする文章分類方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文章分類装置および方法に関し、特に文章の内容に応じて各文章を分類する文章分類装置および方法に関するものである。

【０００２】

【従来の技術】

高度情報化社会では、情報処理技術や情報通信技術の発展に伴い、電子化された膨大な量の情報を容易に入手できる環境が提供されつつある。このような環境を利用して入手した情報は、そのデータ量も膨大となるため、所望する情報を効率よくかつ正確に把握する必要がある。

情報の内容を解析する技術として、各情報を構成する文章の内容に応じて各文章を分類する技術が研究されている。

【０００３】

従来、文章を分類する技術として、予め各分類の内容を示すラベルを用意し、各文章の内容を所定のアルゴリズムで解析し、用意した各ラベルごとにそれぞれの文章を分類するものが提案されている（例えば、非特許文献１など参照）。

これら技術は、文章の分類に際し、予め各分類の内容を示すラベルを用意し、各種の学習アルゴリズムを用いて、これらラベルを各文章に対して精度よく割り当てることにより、各文章をラベルごとに分類しようとするものである。

【０００４】

なお、出願人は、本明細書に記載した先行技術文献情報で特定される先行技術文献以外には、本発明に関連する先行技術文献を出願時までに発見するには至らなかった。

【０００５】

【非特許文献１】

永田昌明他，「テキスト分類－学習論理の見本市－」，情報処理，42巻1号，2001年1月

【非特許文献２】

北研二他,「情報検索アルゴリズム」,共立出版,2002年

【０００６】

【発明が解決しようとする課題】

しかしながら、このような従来の文章分類技術では、予めラベルを用意する必要があるため、分類対象となる各文章の内容をある程度把握して適切なラベルを選択して設定しておく必要がある。したがって、このラベル選択に際し、文章量が多くその内容が広範囲にわたる場合には大きな作業負担を要するという問題点がった。また、分類に用いるラベルは主観的に選択されることから、得られる分類そのものが限定的となり、想定しうる範囲を超えた新たな観点から文章を分類できないという問題点があった。

本発明はこのような課題を解決するためのものであり、比較的少ない作業負担で、主観にとらわれることなく柔軟に分類できる文章分類装置および方法を提供することを目的としている。

【０００７】

【課題を解決するための手段】

このような目的を達成するために、本発明にかかる文章分類装置は、文章集合に含まれる複数の文章を、１つ以上の単語からなるタームを複数有するタームリストに基づき分類する文章分類装置において、各文章と各タームとの関係を２次元表現したＤＴマトリクスを生成するＤＴマトリクス生成手段と、グラフ理論で用いられるＤＭ分解法に基づいて、ＤＴマトリクス生成手段で得られたＤＴマトリクスを変形することにより、変形ＤＴマトリクスを生成するＤＴマトリクス変形手段と、このＤＴマトリクス変形手段で得られた変形ＤＴマトリクス上でブロック化されたクラスタごとに、当該クラスタに属する文章をそれぞれ同一分類として出力する文章分類手段とを備えるものである。

【０００８】

ラベル生成手段をさらに備え、任意のクラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するようにしてもよい。

文章編成手段をさらに備え、変形ＤＴマトリクスでの文章の並び順序に応じて、任意のクラスタに属する文章またはすべての文章を順に出力するようにしてもよい。

要約作成手段をさらに備え、任意の文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力するようにしてもよい。

【０００９】

タームリストの編集に際し、タームリストに対して任意のタームを追加または削除するタームリスト編集手段を設け、指標生成手段で、このタームリスト編集手段による編集前後のタームリストを用いてＤＴマトリクス生成手段によりそれぞれＤＴマトリクスを生成し、これらＤＴマトリクスから当該編集の有用性を示す指標を生成して出力するようにしてもよい。

【００１０】

また、本発明にかかる文章分類方法は、文章集合に含まれる複数の文章を、１つ以上の単語からなるタームを複数有するタームリストに基づき分類する文章分類装置で用いられる文章分類方法において、各文章と各タームとの関係を２次元表現したＤＴマトリクスを生成するステップと、グラフ理論で用いられるＤＭ分解法に基づいてＤＴマトリクスを変形することにより、変形ＤＴマトリクスを生成するステップと、変形ＤＴマトリクス上でブロック化されたクラスタごとに、当該クラスタに属する文章をそれぞれ同一分類として出力するステップとを備えるものである。

【００１１】

この際、任意のクラスタに属する各文章と強連結をなす各タームを、当該クラスタの分類を示すラベルとして出力するステップをさらに備えてもよい。

また、変形ＤＴマトリクスでの文章の並び順序に応じて、任意のクラスタに属する文章またはすべての文章を順に出力するステップをさらに備えてもよい。

また、任意の文章を構成する各文のうち、当該文章と強連結をなすタームを含む文を、当該文章の要約として出力するステップをさらに備えてもよい。

【００１２】

タームリストの編集に際し、タームリストに対して任意のタームを追加または削除するステップと、編集前後のタームリストを用いてそれぞれＤＴマトリクスを生成し、これらＤＴマトリクスから当該編集の有用性を示す指標を生成して出力するステップとをさらに備えてもよい。

【００１３】

【発明の実施の形態】

次に、本発明の実施の形態について図面を参照して説明する。

[文章分類装置の構成]

まず、図１を参照して、本発明の一実施の形態にかかる文章分類装置について説明する。図１は本発明の一実施の形態にかかる文章分類装置の構成を示すブロック図である。

この文章分類装置１は、全体としてコンピュータからなり、制御部１０、記憶部２０、操作入力部３０、画面表示部４０、およびデータ入出力インターフェース部（以下、データ入出力Ｉ／Ｆ部という）５０が設けられている。

【００１４】

制御部１０は、ＣＰＵなどのマイクロプロセッサとその周辺回路からなり、記憶部２０に予め格納されているプログラムを実行して、上記ハードウェアとプログラムとを協働させることにより、文章分類処理のための各種機能手段を実現する。

記憶部２０は、ハードディスクやメモリなどの記憶装置からなり、制御部１０での処理に用いる各種情報を格納する。これら情報としては、分類対象となる各文章からなる文章集合２１や、各文章の内容を把握するための複数の重要語からなるタームリスト２２が記憶されている。

【００１５】

操作入力部３０は、キーボードやマウスなどの入力装置からなり、利用者の操作を検出して制御部１０へ出力する。

画面表示部４０は、ＣＲＴやＬＣＤなどの画面表示装置からなり、制御部１０での処理内容や処理結果を表示出力する。

データ入出力Ｉ／Ｆ部５０は、外部装置（図示せず）や通信ネットワーク（図

示せず）と接続するための回路部であり、文書集合２１やタームリスト２２のほか、得られた処理結果や制御部１０で実行するプログラムをやり取りする際に用いられる。

【００１６】

制御部１０には、機能手段として、ＤＴマトリクス生成手段１１、ＤＴマトリクス変形手段１２、文章分類手段１３、ラベル生成手段１４、文章編成手段１５、要約作成手段１６、タームリスト編集手段１７、タームリスト生成手段１８、および指標生成手段１９が設けられている。

本実施の形態において、ＤＴマトリクスとは、各文章Ｄと各タームＴとの関係を２次元的に表現した行列を指す。

【００１７】

本実施の形態では、上記関係は、文章Ｄ中におけるタームＴの存在有無からなり、文章Ｄとターム Ｔとをそれぞれマトリクスの列と行に対応させ、ある文章Ｄｉがあるターム Ｔｊを含む場合には、ＤＴマトリクスのｊ，ｉ成分を「１」とし、含まない場合には「０」とすることにより、文章Ｄとターム Ｔの関係を表している。

さらに、このＤＴマトリクスを２部グラフの一表現形態と見なし、２部グラフのグラフ理論で用いられるＤＭ分解法に基づきＤＴマトリクスを変形し、得られた変形ＤＴマトリクス上に現れるクラスタに基づき、各文章Ｄを分類するようにしたものである。

【００１８】

ＤＴマトリクス生成手段１１は、分類対象となる各文章Ｄ（Document）とタームリスト２２を構成する各タームＴ（Term）とからＤＴ（Document-Term）マトリクスを生成する機能手段である。

ＤＴマトリクス変形手段１２は、ＤＴマトリクス生成手段１１で生成されたＤＴマトリクスをＤＭ（Dulumage-Mendelsohn）分解法に基づき変形する機能手段である。

【００１９】

ＤＭ分解法とは、具体的には、ＤＴマトリクスに対し、行操作（行同士を入れ

替える操作）または列操作（列同士を入れ替える操作）を施して、三角行列化する処理である。この三角行列化されたＤＴマトリクスを変形ＤＴマトリクスと呼ぶ。

文章分類手段１３は、ＤＴマトリクス変形手段１２で得られた変形ＤＴマトリクス上に現れるブロック化されたクラスタに基づき、文章集合２１の各文章を分類する機能手段である。

【００２０】

ラベル生成手段１４は、各クラスタごとに、当該クラスタに属する各文章Ｄと強連結の関係にあるタームＴを、当該クラスタのラベルとして出力する機能手段である。

文章編成手段１５は、変形ＤＴマトリクスにおける文章Ｄの並び順に基づき、文章集合２１の各文章を並び替えて出力する機能手段である。

要約作成手段１６は、文章Ｄと強連結の関係にあるタームＴを含む文を、当該文章Ｄの要約として出力する機能手段である。

【００２１】

タームリスト編集手段１７は、操作入力部３０からの操作に応じて、記憶部２０のタームリスト２２に対するタームＴの追加／削除を行う機能手段である。

タームリスト生成手段１８は、記憶部２０の文章集合２１に含まれる各文章Ｄを解析して、各文章Ｄの特徴を効果的に表現する語すなわち重要語を抽出し、これら重要語からなるタームＴを用いてタームリスト２２を生成する機能手段である。

指標生成手段１９は、タームリスト編集手段１７で編集されたタームリストについて、その編集前後におけるＤＴマトリクスに基づき当該編集による分類への影響を示す指標を生成する機能手段である。

【００２２】

[文章分類装置の動作]

次に、図２を参照して、本実施の形態にかかる文章分類装置の動作について説明する。図２は本実施の形態にかかる文章分類装置のＤＴマトリクス生成処理を示すフローチャートである。

制御部１０は、操作入力部３０からの指示に応じて、文章分類処理に用いるＤＴマトリクスを生成するため、図２のＤＴマトリクス生成処理を開始する。

まず、ＤＴマトリクス生成手段１１は、記憶部２０に格納されている文章集合２１を読み込むとともに（ステップ１００）、タームリスト２２を読み込む（ステップ１０１）。

【００２３】

図３に文章集合２１の構成例を示す。この例は、「ストレス」についてＷｅｂ上で多数の回答者に自由に文章を記述してもらったものを集計したものであり、各文章Ｄごとに当該文章Ｄを管理するための文章番号Ｄｉとその文章を記述した回答者の識別情報とが割り当てられている。

図４はタームリスト２２の構成例である。このタームリスト２２は、所定のアルゴリズムに基づき各文章Ｄを解析し、得られた重要語の種別とその前後関係とから各タームＴを構成したものであり、各タームＴごとに当該タームＴを管理するターム番号Ｔｊが割り当てられている。

【００２４】

各タームＴは、２つの重要語のうち、前方に位置するキーワード前と後方に位置するキーワード後からなり、それぞれのキーワードごとにそのキーワードの内容を示す単語とその単語の品詞属性種別とが規定されている。また、各タームＴには、後述するタームリスト生成処理により文書集合２１から算出された、文章分類に用いる上での重みを示す重要度が対応付けられている。

例えばターム「１」は、「ストレス」と「解消」という２つのキーワードからなり、その位置関係は「ストレス」が前方に位置するものと規定されている。

ＤＴマトリクス生成手段１１は、文章集合２１内の各文章について、あるしきい値以上の重要度を持った重要度リスト２２の各タームＴが存在するか否かチェックし、その結果からＤＴマトリクスを生成する（ステップ１０２）。

【００２５】

図５にＤＴマトリクスの構成例を示す。このＤＴマトリクス１１Ａは、行方向（縦方向）にタームＴが並べられており、列方向（横方向）に文章Ｄが並べられている。そして、各文章ＤとタームＴの交差位置に、当該文章Ｄにおけるターム

Ｔの存在有無が２進数で記載されている。ここでは、文章Ｄにターム Ｔが存在する場合は「１」が設定され、存在しない場合は「０」が設定されている。

したがって、この例によれば、例えば文章Ｄ１には、タームＴ４，Ｔ７が含まれていることがわかる。またタームＴ２は、文章Ｄ２，Ｄ４に含まれていることがわかる。

【００２６】

続いて、ＤＴマトリクス変形手段１２は、このようにしてＤＴマトリクス生成手段１１で生成されたＤＴマトリクス１１Ａを、ＤＭ分解法に基づき変形して変形ＤＴマトリクス１１Ｂを生成し（ステップ１０３）、これを記憶部２０に格納して、一連のマトリクス生成処理を終了する。

一般に、グラフ理論では、２つの集合に属するそれぞれの点とこれら点を結ぶ辺とからなる２部グラフを、各点間の関連性に基づき分離する手法として、ＤＭ分解法が用いられる。

本実施の形態では、ＤＴマトリクス１１Ａを、文章Ｄからターム Ｔへの辺により結びつけられた２部グラフの一表現形態と見なすことができることに着目し、グラフ理論におけるＤＭ分解法をＤＴマトリクス１１Ａに適用し、得られた変形ＤＴマトリクスに基づき文章Ｄを分類するようにしたものである。

【００２７】

[ＤＭ分解処理]

ここで、図６および図７を参照して、２部グラフにおけるＤＭ分解処理について説明する。図６はＤＭ分解処理を示すフローチャートである。図７はＤＭ分解処の過程を示す２部グラフである。以下では、文章Ｄおよびターム Ｔからなる２つの点集合と、これら点を結ぶ辺からなる２部グラフＧを処理対象とし、これをＤＭ分解法により複数のグラフに分離する場合を例として説明する。なお、これら処理では、制御部１０内部のメモリまたは記憶部２０から各種データを読み出して、制御部１０で所定の演算を行い、その結果を再び記憶するという動作が繰り返し行われる。

【００２８】

まず、図７（ａ）に示すように、処理対象となる２部グラフＧの各辺について

、文章Ｄからターム Ｔへの有向辺を生成する（ステップ２００）。そして、図７（ｂ）に示すように、文章Ｄ側に点ｓを用意し、点ｓから文章Ｄの各点に対して有向辺を生成する（ステップ２０１）。同様にして、ターム Ｔ側に点ｔを用意し、ターム Ｔの各点から点ｔに対して有向辺を生成する（ステップ２０２）。

【００２９】

次に、これら辺を介して点ｓから点ｔへ向かう経路を検索する（ステップ２０３）。例えば図７（ｂ）では、辺２５０，２５１，２５２からなる経路を介して点ｓから点ｔへ向かうことができる。このような経路が存在する場合は（ステップ２０３：ＹＥＳ）、当該経路を構成する各辺を削除するとともに（ステップ２０４）、当該経路上の文章Ｄからターム Ｔへの有向辺とは逆向きの有向辺を、初期状態で空の２部グラフである最大マッチングＭに生成し（ステップ２０５）、ステップ２０３へ戻って次の経路を検索する。図７（ｃ）では、有向辺２５１に対応する逆向きの有向辺２５３が最大マッチングＭに生成されている。

ステップ２０３において、すべての経路の検索が終了して新たな経路が検索されなかった場合（ステップ２０３：ＮＯ）、最大マッチングＭが完成したことになる。

【００３０】

このようにして、図７（ｄ）に示すような最大マッチングＭを完成させた後、最大マッチングＭに属する各有向辺２５４を処理対象Ｇへ含める（ステップ２０６）。これにより、図７（ｅ）に示すように、処理対象Ｇにおいて、最大マッチングＭとして選択された辺２５５については、文章Ｄからターム Ｔへの有向辺とその逆方向の有向辺とから構成されることになる。

次に、ターム Ｔの各点のうち最大マッチングＭに用いられなかった点すなわち自由点２５６を選択し（ステップ２０７）、処理対象Ｇの各辺を介して当該自由点２５６に到達可能な点の集合をクラスタ２６０とする（ステップ２０８）。

【００３１】

同様にして、文章Ｄの各点のうち最大マッチングＭに用いられなかった点すなわち自由点２５７を選択し（ステップ２０９）、処理対象Ｇの各辺を介して当該自由点２５７に到達可能な点の集合をクラスタ２６２とする（ステップ２１０）

。

そして、残りの文章Ｄおよびターム Ｔの各点のうち、双方向に到達可能な経路を有する点集合すなわち強連結をなす点集合をクラスタ２６１とし（ステップ２１１）、一連のＤＭ分解処理を終了する。

このようにして、公知のＤＭ分解法では、各クラスタが所定の順序で生成され、三角行列化された変形ＤＴマトリクスが得られる。

【００３２】

制御部１０では、以上のようにして、図２のＤＴマトリクス生成処理を実行することにより、ＤＴマトリクス生成手段１１で文章集合２１とタームリスト２２とからＤＴマトリクス１１Ａを生成するとともに、ＤＴマトリクス変形手段１２でＤＴマトリクスに対して図６のＤＭ分解処理を適用することにより、各文章Ｄがクラスタごとに分離された変形ＤＴマトリクス１１Ｂを生成する。

【００３３】

図８にＤＴマトリクス１１Ａと変形ＤＴマトリクス１１Ｂの例を示す。ここでは、各文章Ｄｉにおいてターム Ｔｊが存在する場合、その交点にドットが配置されており、ターム Ｔｊが存在しない場合は空白となっている。図８（ａ）のＤＴマトリクス１１Ａでは、ドットがランダムに分布しているが、図８（ｂ）の変形ＤＴマトリクス１１Ｂでは、ドットが断片的ではあるが斜め方向に連続して密集しており、この部分２７０にクラスタが並んでいることがわかる。また、ＤＴマトリクス１１Ｂでは、左下側にドットが存在せず、右上側にドットが多く存在しており、上三角行列化されていることがわかる。

【００３４】

[文章分類処理]

文章分類装置１の制御部１０では、文章集合２１を分類する場合、まず前述のＤＴマトリクス生成処理（図２参照）を実行した後、図９の文章分類処理を実行する。図９は文章分類処理を示すフローチャートである。

まず、文章分類手段１３は、ＤＴマトリクス変形手段１２で生成した変形ＤＴマトリクス１１Ｂ上に現れた各クラスタを識別する（ステップ１１０）。この際、各クラスタについては、変形ＤＴマトリクス１１Ｂを生成した際に分離した部

分グラフに基づき識別してもよく、変形ＤＴマトリクス１１Ｂ上のデータ（ドット）の並びから識別してもよい。

【００３５】

図１０に文章分類処理の説明図を示す。この例では、変形ＤＴマトリクス１１Ｂ上にクラスタ６０が存在している。このクラスタ６０は、２部グラフで表現した場合の部分グラフ６１をなしており、他の文章やタームと関連性が小さい。なお、クラスタ境界が明確な完全グラフをなす場合もある。変形ＤＴマトリクス１１Ｂでは、列方向（横方向）に文章Ｄが並んでおり、クラスタ６０の列方向に並ぶ文章ＤすなわちＤ３６３，Ｄ１５５，Ｄ１５７，Ｄ５，Ｄ１３，Ｄ８が、このクラスタ６０に属する文章Ｄとなる。

文章分類手段１３は、識別された各クラスタに属する各文章からなる部分集合６２を１つの分類として、文章集合２１から抽出して分類し（ステップ１１１）、その結果を例えば画面表示部４０で表示出力し、あるいは記憶部２０へ格納して、一連の文章分類処理を終了する。

【００３６】

このように、本実施の形態では、変形ＤＴマトリクス１１Ｂ上でブロック化されたクラスタごとに、当該クラスタに属する各文章を１つの分類として抽出出力するようにしたので、各分類に対応したラベルを予め用意することなく各文章を分類できる。

したがって、従来のように分類対象となる各文章の内容をある程度把握して適切なラベルを選択する必要がなくなることから、出現頻度など分類に直接関係のない尺度で選択した単語からタームを構成することができ、ラベル選択のための作業負担を大幅に軽減できる。

【００３７】

また、これらクラスタは、複数のタームを橋渡しとして関連付けられた複数の文章から構成されているため、同一タームを含む文章を１つの分類として抽出することができるだけでなく、これら文章内にほぼ共通して存在する他のタームについても、そのタームを含む文章を同一分類として抽出でき、内容に共通性や関連性を持つ文章を１つの分類として容易に抽出できる。

したがって、従来のように予め用意したラベルの有無のみに基づき文章を分類する場合と比較して、そのラベルに限定された主観的な分類ではなく、想定しうる範囲を超えた新たな観点から文章の内容や話題に沿って柔軟に分類を行うことができる。

【００３８】

[ラベル生成処理]

文章分類装置１の制御部１０では、文章分類手段１３で分類された各文章の分類ごとにラベルを生成する場合、まず前述のＤＴマトリクス生成処理（図２参照）および文章分類処理（図９参照）を実行した後、図１１のラベル生成処理を実行する。図１１はラベル生成処理を示すフローチャートである。

まず、ラベル生成手段１４は、ラベルを生成する対象となる分類すなわちクラスタに属する各文章Ｄについて、これら文章Ｄと強連結の関係にあるタームＴを変形ＤＴマトリクス１１Ｂから選択する（ステップ１２０）。

【００３９】

図１２にラベル生成処理の説明図を示す。この例では、任意の分類に属する文章を示す部分集合６２について、各文章Ｄと強連結の関係にあるタームＴ（６３）がそれぞれ選択されている。なお、強連結とは、変形ＤＴマトリクス１１Ｂで各文章Ｄをクラスタごとに分類した際、その２部グラフにおいて、文章Ｄとタームとが互いに双方向の辺で結ばれたペアをいう。通常、これら強連結をなす文章Ｄとタームとは、当該クラスタにおいて対角線上に並ぶ。

次に、選択した各タームＴの単語を当該分類のラベル６４として出力し（ステップ１２１）、その結果を例えば画面表示部４０で表示出力し、あるいは記憶部２０へ格納して、一連のラベル生成処理を終了する。

【００４０】

このように、本実施の形態では、対象となる分類のクラスタに属する各文章と強連結の関係にあるタームＴを、当該分類のラベルとして出力するようにしたので、本実施の形態のように予め用意されたラベルに基づき文章を分類するものではない場合でも、各分類の特徴を単語で表現した適切なラベルを容易に生成できる。

【００４１】

[文章編成処理]

文章分類装置１の制御部１０では、各文章Ｄの並びを編成する場合、まず前述のＤＴマトリクス生成処理（図２参照）を実行した後、図１３の文章編成処理を実行する。図１３は文章編成処理を示すフローチャートである。

まず、文章編成手段１５は、変形ＤＴマトリクス１１Ｂ上での並びに基づき、各文章Ｄを並び替える（ステップ１３０）。

【００４２】

図１４に文章編成処理の説明図を示す。前述したように、ＤＴマトリクスをＤＭ分解法により変形して得られた変形ＤＴマトリクス１１Ｂにおいて、各文章Ｄはターム Ｔを仲立ちとして互いに関連性の高いものが隣接して並んでいる。

文章編成手段１５は、このような変形ＤＴマトリクス１１Ｂに基づき並び変えられた文章Ｄを編成し、編成された各文章６５を出力し（ステップ１３１）、その結果を例えば画面表示部４０で表示出力し、あるいは記憶部２０へ格納して、一連の文章編成処理を終了する。

【００４３】

特に、変形ＤＴマトリクス１１Ｂには、文章Ｄおよびターム Ｔの並びに所定の半順序が存在する。ＤＴマトリクス１１Ａは、ターム Ｔを変数とする文章Ｄの線形連立方程式を示す行列と見なすことができ、変形ＤＴマトリクス１１Ｂは、これら各方程式の解が求まる順序にほぼ沿った順序で文章Ｄが並び替えられた結果を示している。このことからも、変形ＤＴマトリクス１１Ｂ上の文章Ｄの並びには、前後の文章Ｄとの関連性が高いことがわかる。

【００４４】

このように、本実施の形態では、変形ＤＴマトリクス上の文章Ｄの並びに基づき、各文章Ｄを並び替えて出力するようにしたので、共通のタームすなわち単語を持った関連性の高い文章が順に得られることになり、前後の文章Ｄと話題の共通性が得られる。したがって、内容が類似した文章が前後に並べられていることから、アトランダムに文章Ｄを読む場合と比較して、文脈が途切れることなく読むことができクラスタさらには文書集合全体の内容を容易に把握できる。

この際、任意のクラスタすなわち分類に含まれる各文章Ｄを文書編成の対象として１つの文書を生成してもよく、文章集合２１に含まれるすべての文章Ｄを文章編成の対象として１つの文書を生成してもよい。

【００４５】

[要約作成処理]

文章分類装置１の制御部１０では、複数の文からなる任意の文章Ｄの要約を作成する場合、まず前述のＤＴマトリクス生成処理（図２参照）を実行した後、図１５の要約作成処理を実行する。図１５は要約作成処理を示すフローチャートである。

まず、要約作成手段１６は、対象となる文章Ｄについて、前述したラベル生成処理と同様にして、その文章Ｄと強連結の関係にあるタームＴを変形ＤＴマトリクス１１Ｂから選択する（ステップ１４０）。

【００４６】

図１６に要約作成処理の説明図を示す。通常、文章Ｄ（６６）は、複数の文から構成されており、これら文のいずれかに文章Ｄと強連結のタームＴ（６７）が含まれていることになる。この際、このタームＴは文章Ｄの特徴を示していることになる。

要約作成手段１６は、このタームＴを含む文を当該文章Ｄから選択して、これら文を当該文章Ｄの要約６８として出力し（ステップ１４１）、その結果を例えば画面表示部４０で表示出力し、あるいは記憶部２０へ格納して、一連の要約作成処理を終了する。

【００４７】

このように、本実施の形態によれば、対象となる文章Ｄと強連結の関係にあるタームＴに基づいて、そのタームを含む文を当該文章Ｄの要約として出力するようにしたので、文章Ｄの要約を極めて容易にかつ適切に作成できる。

【００４８】

[タームリスト生成処理]

タームリスト生成手段１８は、文章集合２１からタームリスト２２を自動生成するものである。

文章からその文章を特徴付ける重要語を抽出する方法として、各種のアルゴリズムが提案されている。例えば、各単語の重要度を算出し、その重要度に基づき重要語を選択するＴＦＩＤＦ（Term Frequency Inverce Document Frequency）などのアルゴリズムを用いてもよい。あるいは、言語学的な解釈に基づかないフレーズ（共起語）を、辞書を用いることなく抽出するＫｅｙＧｒａｐｈというアルゴリズムを用いてもよい（例えば、非特許文献２など参照）。

【００４９】

タームリスト生成手段１８では、このような公知のアルゴリズムを用いてタームリスト２２を生成する。本実施の形態では、これら単語を特定するため、各単語の品詞属性を形態素解析により予め求めておき、単語のとの品詞属性をペアとして重要語を構成している。また、本実施の形態では、２つの重要語の出現順序を規定したものをタームとして定義しており、これにより文章の内容をより適切にタームで表現可能となっている。

なお、このタームリスト２２については、タームリスト編集手段１７で、操作入力部３０からの指示に基づき生成してもよく、データ入出力Ｉ／Ｆ部５０を介して予め用意されたものを装置外部から入力するようにしてもよい。

【００５０】

[指標生成処理]

タームリスト２２は、変形ＤＴマトリクス１１Ｂを生成して文章を分類する上で重要なファクタとなることから、重要語リスト編集手段１７で、このタームリストを編集可能としている。

本実施の形態では、編集されたタームリストについて、制御部１０の指標生成手段１９により客観的な評価値を算出し、その編集に対する指標を生成する。以下、図１７を参照して、指標生成手段１９における指標生成処理について説明する。図１７は指標生成処理を示すフローチャートである。

【００５１】

まず、タームリスト編集手段１７により、タームリスト２２についてタームＴｋを追加または削除し、新たなタームリストが生成されたものとする（ステップ１５０）。指標生成手段１９では、編集前後のタームリストのそれぞれについて

、ＤＴマトリクス生成手段１１によりＤＴマトリクスを生成し（ステップ１５１）、各ＤＴマトリクスごとに平均文章類似度Ｑを算出する（ステップ１５２）。

平均文章類似度Ｑは、２つの文章Ｄｉ，Ｄｊ間の類似度ｓｉｍ（Ｄｉ，Ｄｊ）をすべての文章間について算出し平均したものであり、文章Ｄの数をＮとした場合、Ｑは次の数１で算出される。

【００５２】

【数１】

$$Q=\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}sim(Di,Dj)\cdots(1)$$

【００５３】

この際、類似度ｓｉｍ（Ｄｉ，Ｄｊ）は、当該変形ＤＴマトリクスに基づき、文章Ｄｉ，Ｄｊにおける各タームＴの有無を０／１で示すベクトルをＸ，Ｙとした場合、例えば数２～数４により算出される。特に、数２はベクトルＸ，Ｙの内積を類似度とするもの、数３はベクトルＸ，ＹのＤｉｃｅ係数を類似度とするもの、数４はベクトルＸ，ＹのＪａｃｃａｒｄ係数を類似度とするものである。

【００５４】

【数２】

$$sim(Di,Dj)=|X\cap Y|=\sum_{i=1}^{t}xi\cdot yi\cdots(2)$$

【００５５】

【数３】

$$sim(Di,Dj)=\frac{2|X\cap Y|}{|X|+|Y|}\cdots(3)$$

【００５６】

【数４】

$$sim(Di,Dj)=\frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|}\cdots(4)$$

【００５７】

このようにして、指標生成手段１９は、編集前のタームリストから生成された

ＤＴマトリクスに基づき平均文章類似度Ｑを算出するとともに、編集後のタームリストから生成されたＤＴマトリクスに基づき平均文章類似度Ｑｋを算出して、これらの差ΔＱ＝Ｑｋ－Ｑを算出し、指標値として画面表示部４０から表示出力する（ステップ１５３）。

ここで、差ΔＱがゼロより大きい場合は（ステップ１５４：ＹＥＳ）、編集後のタームリストから生成されたＤＴマトリクスのほうが、各文章の類似度が大きくなり、各文章が効果的に分類できることから、当該編集は有効である旨を画面表示部４０へ表示出力し（ステップ１５５）、一連の指標生成処理を終了する。

【００５８】

また、ステップ１５４において、差ΔＱがゼロ以下の場合は（ステップ１５４：ＮＯ）、編集後のタームリストから生成されたＤＴマトリクスのほうが、各文章の類似度が小さくなり、各文章が効果的に分類できないことから、当該編集は無効である旨を画面表示部４０へ表示出力し（ステップ１５６）、一連の指標生成処理を終了する。

なお、指標としては、ΔＱだけを表示出力して作業者に編集の有効性を判断させるようにしてもよい。また当該編集に対する有効／無効だけを表示出力してもよい。

【００５９】

このように、本実施の形態では、指標生成手段１９により、編集前後のタームリストから生成されたＤＴマトリクスに基づき平均文章類似度Ｑを算出し、その変化により当該編集の有効性を示す指標を生成するようにしたので、タームリスト２２に対する編集の有効性を容易に把握することができる。したがって、容易かつ適切にタームリストを編集でき、この編集により所望の意図や目的に応じて効率よく文章を分類することができる。

また、ＤＴマトリクスから得られた平均文章類似度に基づき指標を生成するようにしたので、文章を分類する必要がなくなり指標生成に要する処理を簡素化できる。したがって、当該編集に対する有効／無効を迅速に判断でき、タームリストの編集に要する作業負担を大幅に軽減できる。

【００６０】

なお、平均文章類似度Ｑを用いて当該編集に対する有効／無効を判断する場合について説明したが、これに限定されるものではない。例えば文章を分類した結果、例えば分類数や１分類に属する文章数などに基づき当該編集に対する有効／無効を判断するようにしてもよい。

【００６１】

【発明の効果】

以上説明したように、本発明は、文章集合内の各文章とタームリスト内の各タームとからＤＴマトリクスを生成した後、そのＤＴマトリクスをＤＭ分解し、得られた変形ＤＴマトリクス上の各クラスタごとに、当該クラスタに属する各文章を１つの分類として抽出出力するようにしたので、各分類に対応したラベルを予め用意することなく各文章を分類できる。

したがって、従来のように分類対象となる各文章の内容をある程度把握して適切なラベルを選択する必要がなくなることから、出現頻度など分類に直接関係のない尺度で選択した単語からタームを構成することができ、ラベル選択のための作業負担を大幅に軽減できる。

【００６２】

また、これらクラスタは、複数のタームを橋渡しとして関連付けられた複数の文章から構成されているため、同一タームを含む文章を１つの分類として抽出することができるだけでなく、これら文章内にほぼ共通して存在する他のタームについても、そのタームを含む文章を同一分類として抽出でき、内容に共通性や関連性のある文章を１つの分類として容易に抽出できる。

したがって、従来のように予め用意したラベルの有無のみに基づき文章を分類する場合と比較して、そのラベルに限定された主観的な分類ではなく、想定しうる範囲を超えた新たな観点から文章の内容や話題に沿って柔軟に分類を行うことができる。

【図面の簡単な説明】

【図１】　本発明の一実施の形態にかかる文章分類装置の構成を示すブロック図である。

【図２】　ＤＴマトリクス生成処理を示すフローチャートである。

【図３】　文章集合の構成例である。

【図４】　タームリストの構成例である。

【図５】　ＤＴマトリクスの構成例である。

【図６】　ＤＭ分解処理を示すフローチャートである。

【図７】　ＤＭ分解処理の過程を示す２部グラフである。

【図８】　ＤＴマトリクスおよび変形ＤＴマトリクスの例である。

【図９】　文章分類処理を示すフローチャートである。

【図１０】　文章分類処理を示す説明図である。

【図１１】　ラベル生成処理を示すフローチャートである。

【図１２】　ラベル生成処理を示す説明図である。

【図１３】　文章編成処理を示すフローチャートである。

【図１４】　文章編成処理を示す説明図である。

【図１５】　要約作成処理を示すフローチャートである。

【図１６】　要約作成処理を示す説明図である。

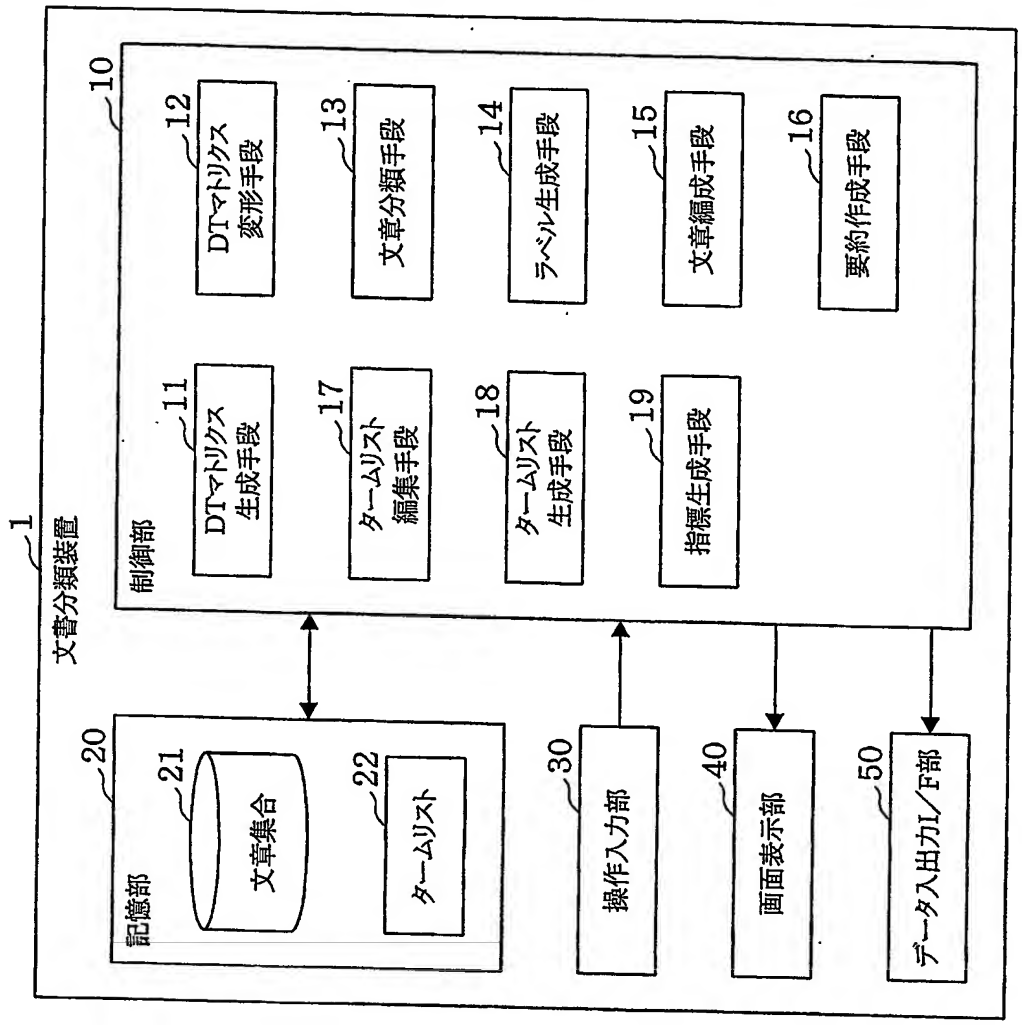
【図１７】　指標生成処理を示すフローチャートである。

【符号の説明】

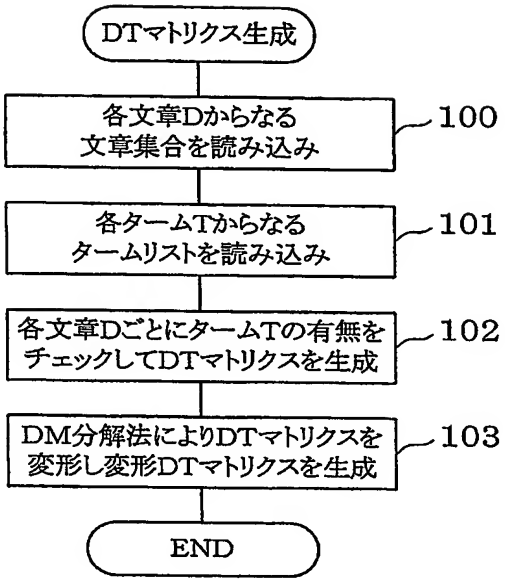
１…文章分類装置、１０…制御部、１１…ＤＴマトリクス生成手段、１１Ａ…ＤＴマトリクス、１１Ｂ…変形ＤＴマトリクス、１２…ＤＴマトリクス変形手段、１３…文章分類手段、１４…ラベル生成手段、１５…文章編成手段、１６…要約作成手段、１７…タームリスト編集手段、１８…タームリスト生成手段、１９…指標生成手段、２０…記憶部、２１…文章集合、２２…タームリスト、３０…操作入力部、４０…画面表示部、５０…データ入出力Ｉ／Ｆ部、６０…クラスタ、６１…部分グラフ、６２…部分集合（分類文章）、６３…強連結ターム、６４…ラベル、６５…編成された文章、６６…文章、６７…強連結ターム、６８…要約。

【書類名】　図面

【図１】


1
文書分類装置
10
制御部
11
DTマトリクス生成手段
12
DTマトリクス変形手段
13
文章分類手段
14
ラベル生成手段
15
文章編成手段
16
要約作成手段
17
タームリスト編集手段
18
タームリスト生成手段
19
指標生成手段
20
記憶部
21
文章集合
22
タームリスト
30
操作入力部
40
画面表示部
50
データ入出力I/F部

【図２】



【図３】

文章集合　21

DI	回答者	文章内容
1	W0039411	カラオケも楽しいですね。リズムにのって歌えるし間違っても誰も止めないですもんね。団でうたうのは周りの目も厳しいし頑張ら…
351	M0014787	現実逃避？なのかも知れませんが、私はストレスを感じたときは、ひたすら眠るようにしています。いい夢を見れば気分爽快なん…
289	M0010523	＞好きな香りの入浴剤を入れて、のんびり入浴。＞入浴しながら読書するとストレス解消になります。ぼくも本を持ち込んで読ん…
319	W0013732	マッサージって気持ちいいけど、ツボを心得ていない人にされるとかえってストレスになったりするのよね。マッサージも行きつけ…
57	W0039210	いいですね。気持ちよさそう!!! 私のように全然ばっとに当たらなければ、逆にストレスになってしまうかも… こういうストレスの…
72	W0039200	私もプールに通っています。ただがんばりすぎると逆に疲れるので三十分程度にして好きなように泳いだり歩いたりしています。…
111	W0012712	先日たった一泊ですが何年ぶりかで旅行に行ってきました。食事の準備や布団の上げ下ろしもしなくていいのがこんな幸せな…
337	W0013147	＞私のストレス解消法はなんといってもカーデニング！＞旦那も子供も文句ばかり言うけど、花は何の文句も言わずに咲いてく…
360	W0015958	3歳になる息子がいます。公園に遊びに行ったりしますが、大の大人ならブランコなんか恥ずかしくて乗れませんが、子供と一緒…
32	W0016759	私も不良主婦なんです。胸を張って家事は完璧ですなんて言えないのに、2～3ヶ月に1回くらいの割で夜遊びします。7時ごろ…
		⋮

【図４】

タームリスト　22

TJ	キーワード前		キーワード後		重要度
1	ストレス	名詞－一般	解消	名詞－サ変接続	256
2	解消	名詞－サ変接続	ストレス	名詞－一般	256
3	ストレス	名詞－一般	仕事	名詞－サ変接続	117
4	仕事	名詞－サ変接続	ストレス	名詞－一般	117
5	とき	名詞－非自立－副詞可能	ストレス	名詞－一般	116
6	吸う	動詞－自立	ストレス	名詞－一般	99
7	的	名詞－接尾－形容動詞語幹	ストレス	名詞－一般	88
8	行く	動詞－自立	ストレス	名詞－一般	86
9	寝酒	名詞－一般	晩酌	名詞－一般	85
10	晩酌	名詞－一般	寝酒	名詞－一般	85
11	人	名詞－接尾－助数詞	ストレス	名詞－一般	83
12	子供	名詞－一般	ストレス	名詞－一般	77
			⋮		

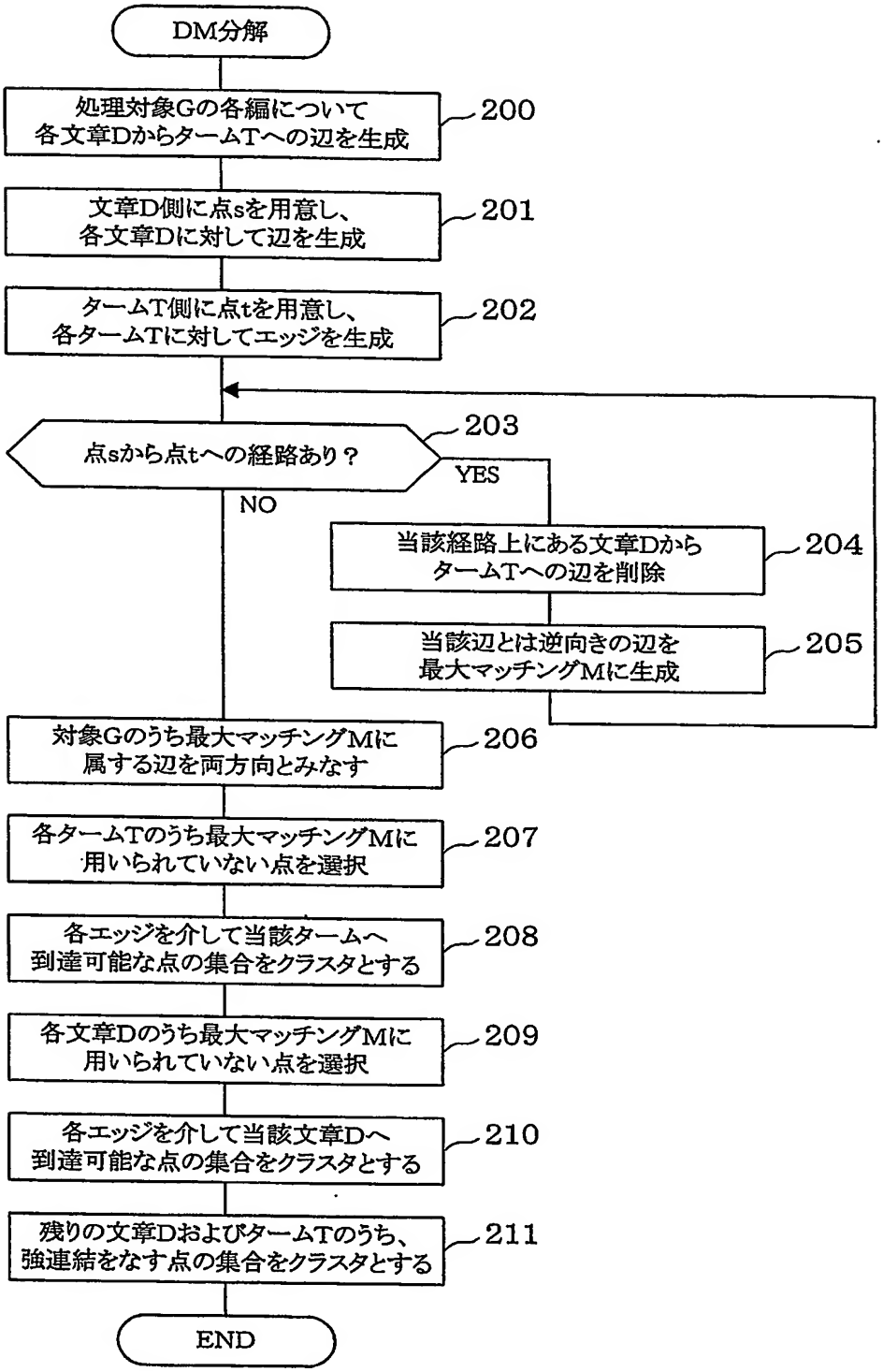
【図５】

DTマトリクス　11A

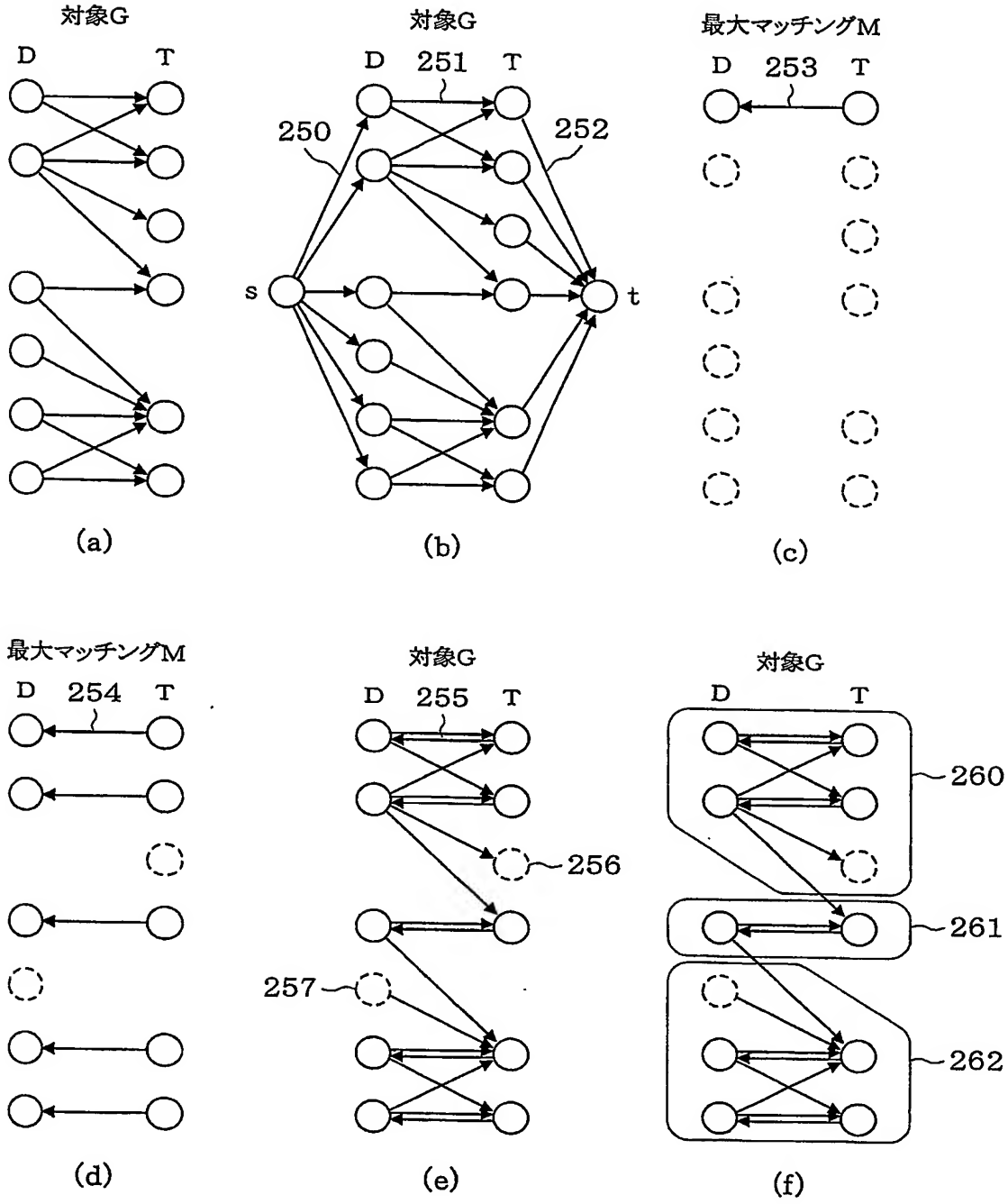
		文章D							
		D1	D2	D3	D4	D5	D6	⋯⋯	Dm
タームT	T1	0	1	0	0	0	0		0
	T2	0	1	0	1	0	0		0
	T3	0	0	0	0	0	1		0
	T4	1	0	0	0	0	0		0
	T5	0	0	0	0	0	0	⋯⋯	1
	T6	0	0	0	0	1	0		1
	T7	1	0	0	0	0	0		0
	T8	0	0	0	0	0	0		0
	⋮	⋮		⋮				⋱	⋮
	Tn	0	0	0	1	0	0	⋯⋯	0

0＝文章Diにターム TJが存在しない
1＝文章Diにターム TJが存在する

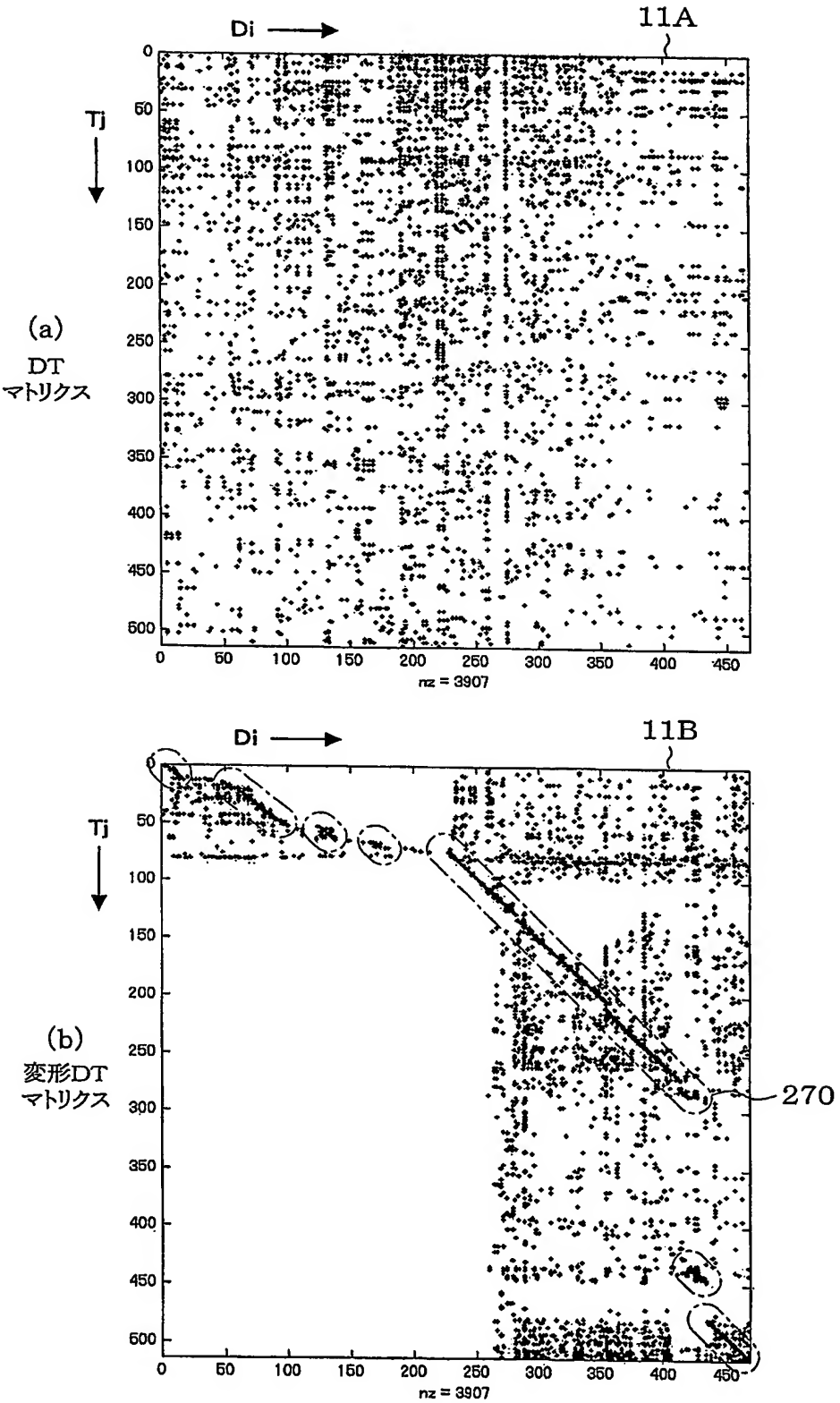
【図６】


DM分解
処理対象Gの各編について
各文章DからタームTへの辺を生成
200
文章D側に点sを用意し、
各文章Dに対して辺を生成
201
タームT側に点tを用意し、
各タームTに対してエッジを生成
202
点sから点tへの経路あり？
203
YES
NO
当該経路上にある文章Dから
タームTへの辺を削除
204
当該辺とは逆向きの辺を
最大マッチングMに生成
205
対象Gのうち最大マッチングMに
属する辺を両方向とみなす
206
各タームTのうち最大マッチングMに
用いられていない点を選択
207
各エッジを介して当該タームへ
到達可能な点の集合をクラスタとする
208
各文章Dのうち最大マッチングMに
用いられていない点を選択
209
各エッジを介して当該文章Dへ
到達可能な点の集合をクラスタとする
210
残りの文章DおよびタームTのうち、
強連結をなす点の集合をクラスタとする
211
END

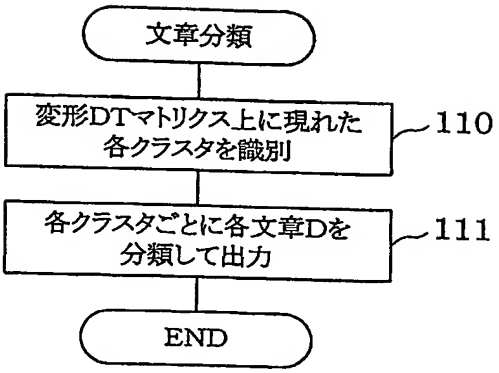
【図７】


対象G
D
T
(a)
対象G
D
251
T
250
252
s
t
(b)
最大マッチングM
D
253
T
(c)
最大マッチングM
D
254
T
(d)
対象G
D
255
T
256
257
(e)
対象G
D
T
260
261
262
(f)

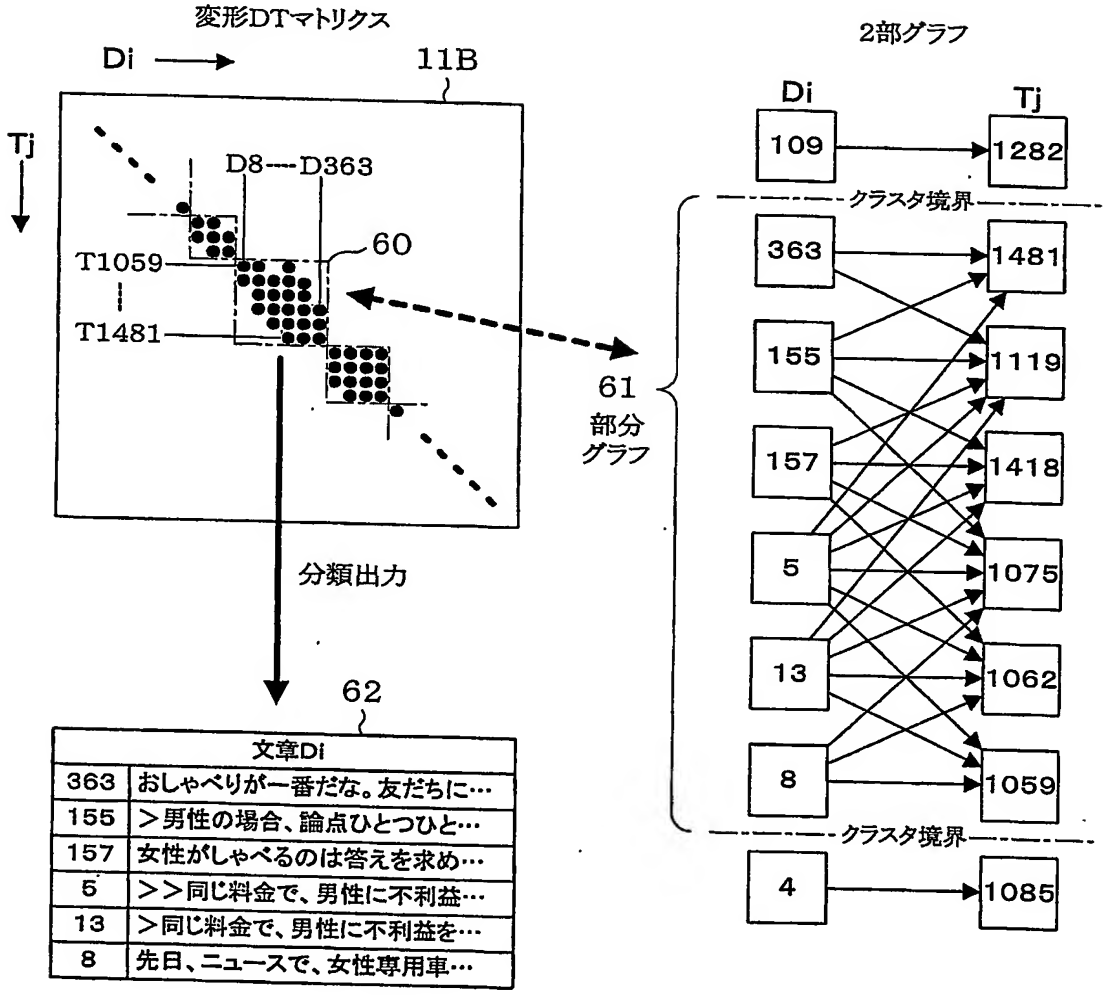
【図８】


Di
11A
Tj
(a)
DT
マトリクス
0
50
100
150
200
250
300
350
400
450
500
nz = 3907
Di
11B
Tj
(b)
変形DT
マトリクス
270
nz = 3907

【図９】

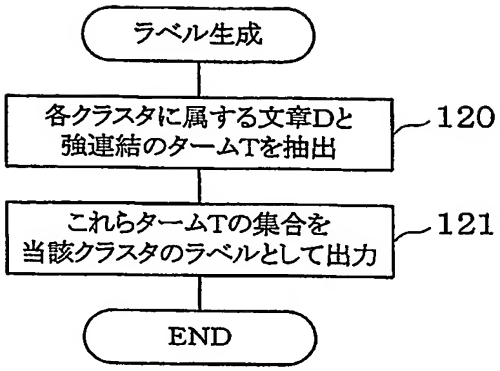


【図１０】

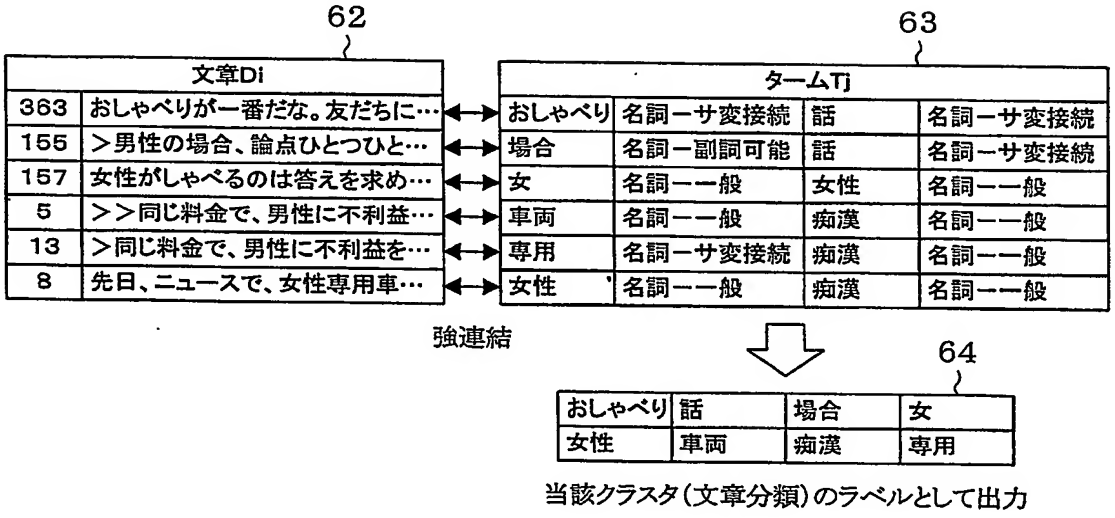


文章Di	
363	おしゃべりが一番だな。友だちに…
155	＞男性の場合、論点ひとつひと…
157	女性がしゃべるのは答えを求め…
5	＞＞同じ料金で、男性に不利益…
13	＞同じ料金で、男性に不利益を…
8	先日、ニュースで、女性専用車…

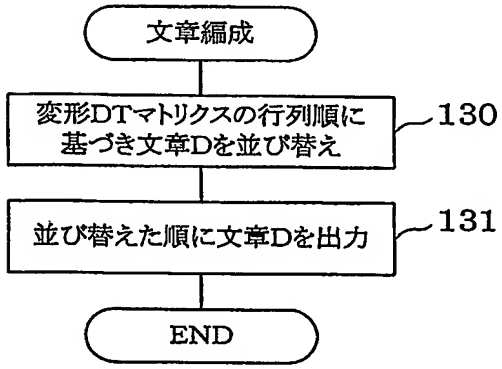
【図１１】


ラベル生成
各クラスタに属する文章Dと
強連結のタームTを抽出
120
これらタームTの集合を
当該クラスタのラベルとして出力
121
END

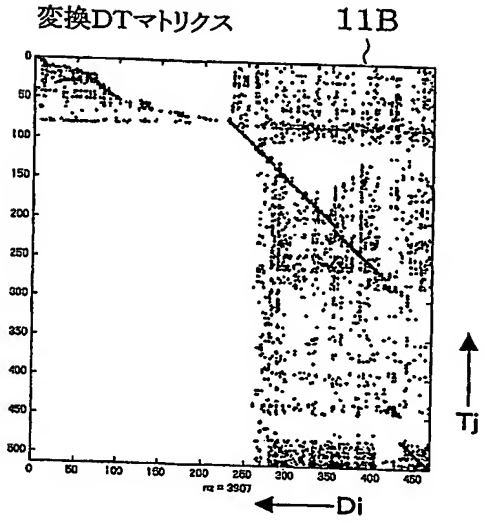
【図１２】

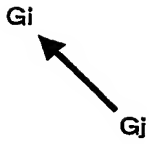

62
63
文章Di
363 おしゃべりが一番だな。友だちに…
155 ＞男性の場合、論点ひとつひと…
157 女性がしゃべるのは答えを求め…
5 ＞＞同じ料金で、男性に不利益…
13 ＞同じ料金で、男性に不利益を…
8 先日、ニュースで、女性専用車…
タームTj
おしゃべり 名詞ーサ変接続 話 名詞ーサ変接続
場合 名詞ー副詞可能 話 名詞ーサ変接続
女 名詞ー一般 女性 名詞ー一般
車両 名詞ー一般 痴漢 名詞ー一般
専用 名詞ーサ変接続 痴漢 名詞ー一般
女性 名詞ー一般 痴漢 名詞ー一般
強連結
64
おしゃべり 話 場合 女
女性 車両 痴漢 専用
当該クラスタ（文章分類）のラベルとして出力

【図１３】


文章編成
変形DTマトリクスの行列順に
基づき文章Dを並び替え
130
並び替えた順に文章Dを出力
131
END

【図１４】





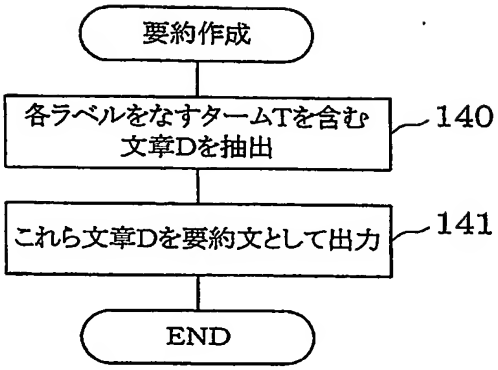
線形連立方程式の解が
求まる順序として、
変形DTマトリクス上に
半順序Gi≦Gjが存在

変形DTマトリクスにおける文章Dの並びに基づき各文書Dを並び替え

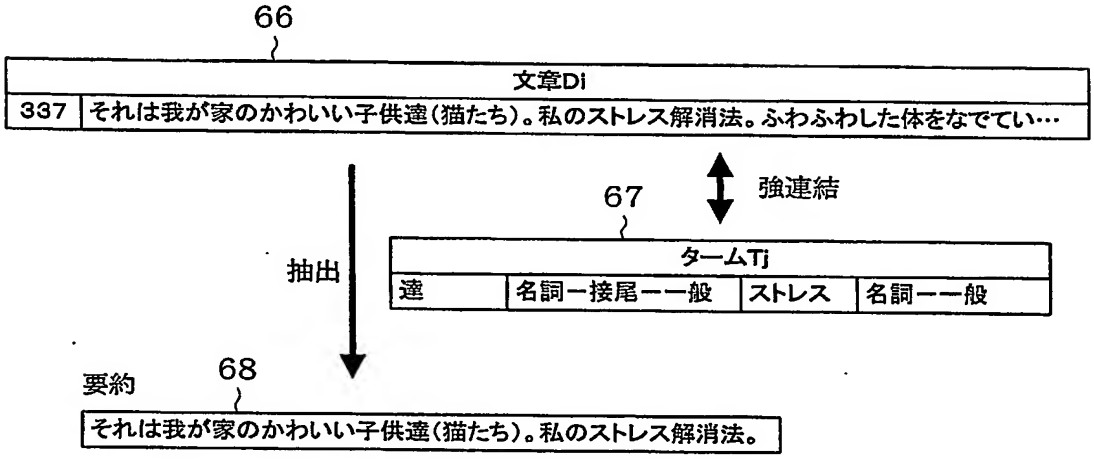
65

	文章Di
1	私はあんまりストレスを感じませんが… みんなそれぞれ違いますしストレスを感じてしまう…
351	現実逃避？なのかもれませんが、私はストレスを感じたときはひたすら眠るようにしています。
289	ぼくも本を持ち込んで読んだことがあるんですが、あがったときに本がブヨブヨ(笑)なんとか…
319	(笑)。マッサージって気持ちいいけど、ツボを心得ていない人にされるとかえってストレスに…
57	気持ちよさそう!!! 私のように全然バットに当たらなければ、逆にストレスになってしまうかも…
72	ただがんばりすぎると逆に疲れるので三十分程度にして好きなように泳いだり歩いたりして…
111	温泉にゆっくり入ってストレスも吹っ飛んでしまいました。
337	＞私のストレス解消法はなんといってもガーデニング！＞旦那も子供も文句ばっかり言うけ…
360	公園に遊びに行ったりしますが 大の大人なら ブランコなんか恥ずかしくて乗れませんが…
32	メッチャ楽しい！メンバーは子供会活動を一緒にやった人達8名。
390	子供の頃、銭湯に行ったことがあったけど大きくて泳げたりしたのがたのしかったです。
227	雑誌なんかを読みながら＞トリートメントをしたり、ボディマッサージなんかして＞なんか自…
	⋮

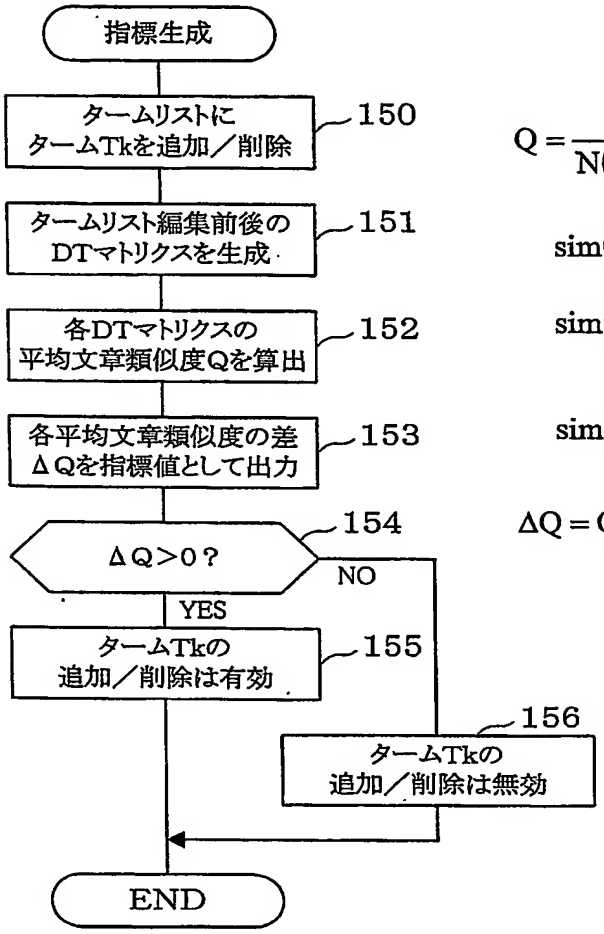
【図１５】


要約作成
各ラベルをなすタームTを含む文章Dを抽出
140
これら文章Dを要約文として出力
141
END

【図１６】


66
文章Di
337
それは我が家のかわいい子供達（猫たち）。私のストレス解消法。ふわふわした体をなでてい…
67
強連結
抽出
タームTj
達
名詞－接尾－一般
ストレス
名詞－一般
要約
68
それは我が家のかわいい子供達（猫たち）。私のストレス解消法。

【図１７】



$$Q=\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\mathrm{sim}(Di,Dj)\cdots(1)$$

$$\mathrm{sim}(Di,Dj)=|X\cap Y|=\sum_{i=1}^{t}xi\cdot yi\cdots(2)$$

$$\mathrm{sim}(Di,Dj)=\frac{2|X\cap Y|}{|X|+|Y|}\cdots(3)$$

$$\mathrm{sim}(Di,Dj)=\frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|}\cdots(4)$$

$$\Delta Q=Qk-Q\cdots(5)$$

【書類名】　　要約書

【要約】

【課題】　比較的少ない作業負担で、主観にとらわれることなく柔軟に分類できるようにする。

【解決手段】　ＤＴマトリクス生成手段１１で、文章集合２１内の各文章Ｄとタームリスト２２内の各タームＴとからＤＴマトリクス１１Ａを生成し、ＤＴマトリクス変形手段１２で、そのＤＴマトリクス１１ＡをＤＭ分解して変形ＤＴマトリクス１１Ｂを生成する。そして文章分類手段１３で、変形ＤＴマトリクス１１Ｂ上に現れる各クラスタごとに、当該クラスタに属する各文章Ｄを１つの分類（部分集合）として抽出出力する。

【選択図】　　図１

特願２００３－１８９７１６

出願人履歴情報

識別番号　［０００００６６６６］

１．変更年月日　１９９８年　７月　１日
［変更理由］　名称変更
住　所　東京都渋谷区渋谷２丁目１２番１９号
氏　名　株式会社山武

出証番号　出証特２００４－３０７１９７５